Sample Size/Power

Michael Proschan, Ph.D.
Mathematical Statistician
National Heart, Lung, and Blood Institute

Outline

- Introduction
- Continuous Outcome Trials
  - Nuisance parameter: Standard deviation
  - Treatment effect: Difference in means
- Dichotomous Outcome Trials
  - Nuisance parameter: Control event rate
  - Treatment effect: Difference in proportions
- Adaptive Sample Size Methods

- A clinical trial is set up to make it difficult to reject H_0 and declare the treatment beneficial
- Because the type I error rate is small, if we declare treatment beneficial, it probably is
- But what if we don’t declare the treatment beneficial? Can we be confident that it isn’t beneficial?
- Only if power is high!

- *Power* is probability of correctly declaring treatment benefit
- If power is high and you don’t find a treatment difference, you can be confident of no real difference (if there were a real difference, you likely would have seen it)

- Power depends on size of treatment effect δ
  - In t-test comparing blood pressure change,

    δ=(mean BP change)$_T$-(mean BP change)$_C$ in population of millions
  - In comparison of proportions who quit smoking,

    δ=(proportion quitting)$_T$-(proportion quitting)$_C$ in population of millions
- The larger the treatment effect in population, the greater the power of trial

- Power and the 3 bears:
  - Mama Bear's power was too low, so she missed a large treatment effect
  - Papa Bear's power was too high, so he wasted resources and declared a tiny, clinically irrelevant effect statistically significant
  - Baby Bear's power was just right, so he had a good chance of detecting reasonably sized effects, but not tiny ones

- Unfortunately, power depends on *nuisance parameters* as well (population quantities of little or no intrinsic interest, but are needed to compute power)
  - In t-test, we need to estimate the standard deviation σ
  - In proportions test, we need to estimate the control rate p_C

- The nuisance parameter is big nuisance!
  - If standard deviation estimate is too small or control event rate estimate too large, trial is underpowered
  - If standard deviation estimate is too large or control event rate estimate too low, the trial is larger than it needs to be

Continuous Outcomes

- For continuous outcome trial using α=.05 two-tailed t-test, per-arm sample size for 90% power is

$$n=2\sigma^2(1.96+1.28)^2/\delta^2$$

- σ=standard deviation and δ=treatment effect
- For 80% or 85% power, replace 1.28 by .84 or 1.04, respectively

Example

- E.g., compare yoga to meditation to lower blood pressure
- Outcome: change in systolic blood pressure (SBP) from baseline to 6 weeks
- To determine sample size, need to specify standard deviation σ (nuisance) and treatment effect δ

Example: Specifying σ

- Note: Outcome is baseline to 6 week change, so standard deviation must be standard deviation of baseline to 6 week change
- How do we estimate σ?
- Better to overestimate than underestimate
- Best strategy: Estimate from similar trial

Example: Specifying σ

- What if similar trial lasts 3 weeks? Use standard deviation of baseline to 3 week change to estimate standard deviation of baseline to 6 week change?
- No! The longer the duration, the larger the standard deviation of change; using baseline to 3 week change will underestimate σ
- If similar trial used baseline to 12 week change, get conservative estimate of σ (good)

Example: Specifying σ

- What if no similar trials, but have epidemiology study of changes in SBP over time?
- Use standard deviation of baseline to 6 week change in epi study?
- Increase it! Interventions can increase standard deviation (intervention may help some people, have no effect on others)

Example: Specifying σ

- What if have no info on standard deviation of baseline to 6 week change? All we know is standard deviation of SBPs at single time is 11 mm Hg
- Useful formula relating standard deviation of change to standard deviation at single time and correlations:

Example: Specifying σ

std dev(change)=$\{2(1-\rho)\}^{1/2}$ std dev(single)

where ρ is correlation between baseline and end of study measurements

If know correlation and standard deviation at single time point, can get std dev of change

Example: Specifying σ

- The shorter the trial, the higher ρ is (with 6 week BP trial, correlation is around .90)
- Be conservative: Err on side of underestimating ρ
- With ρ=.8 and std deviation at single time 11, std deviation of change is estimated to be $\{2(1-.8)\}^{1/2}(11)\approx 7$

Example: Specifying δ

- Two approaches to specifying δ:
  - Determine smallest relevant effect or
  - Look at effects seen in similar clinical trials

- Approach 1: Not testing medicine with side effects; from public health standpoint, even small blood pressure differences (δ=2 mm Hg) worth detecting

Example: Specifying δ

- Approach 2: Look at similar trials, if any
- Maybe most similar trial compares meditation to no meditation, & found 4 mm Hg difference
- Now must decide whether to size for $\delta=2$, $\delta=4$, or something between 2 and 4
- Decision may depend on sample size!

Example: Specifying δ

- Suppose use standard deviation of change 7 mm Hg
- For 90% power to detect 2 mm Hg difference, need

$$n=2(7^2)(1.96+1.28)^2/2^2=258/\text{arm}$$

- Need 2(258)=516 people

Example: Specifying δ

- To detect 4 mm Hg difference, need

$$n=2(7^2)(1.96+1.28)^2/4^2=65/arm$$

- Now only need 2(65)=130 people
- Doubling the effect decreases sample size four-fold!

Example: Specifying δ

- Look at sample sizes for other treatment effects, e.g., for $\delta=3$ and 90% power,

  $n=2(7^2)(1.96+1.28)^2/3^2=115$/arm

  Make a decision based on detectable effect and sample size

- Note: δ is expected *net* treatment effect, taking into account that some people assigned to yoga won't do it, and some assigned to meditation may do yoga
- If expect $\delta=4$ under perfect compliance, want to use smaller number to account for noncompliance

Fixed Total Sample Size

- Sometimes sample size is fixed (can afford 100/arm) & want to see what it buys you
- Power with n/arm and 2-tailed α=.05 is

$$\text{Power}=\Phi\{\delta/(2\sigma^2/n)^{1/2}-1.96\}$$

where Φ is the standard normal distribution function

Fixed Total Sample Size

- E.g., suppose we can afford only 100/arm
- Power to detect a 2 mm Hg difference is

$$\text{Power}=\Phi[2/\{2(7)^2/100\}^{1/2}-1.96]$$

$$=\Phi\{0.06)=.52$$

- Only 52% power to detect a 2 mm Hg difference

Fixed Total Sample Size

- Could also determine effect detectable with 90% power with sample size of 100/arm:

$$\delta=(1.96+1.28)(2\sigma^2/n)^{1/2}$$

$$=3.24\{2(7)^2/100\}^{1/2}$$

$$=3.21 \text{ mm Hg}$$

- 90% power for 3.21 mm Hg difference

Sensitivity Analysis

- Power depends on both the treatment effect and standard deviation
- You should always do sensitivity analyses to see how power changes for different values of σ and δ
- Make table:

Power with 258/arm

	δ=1.5	δ=2	δ=2.5
σ=6	.81	.97	>.99
σ=7	.68	.90	.98
σ=8	.57	.81	.94

Common Errors in Interpretation of Power

- 90% power to detect a 2 mm Hg reduction
  - E1) “If *observed* reduction < 2 mm Hg, we won’t get a statistically significant result”
  - E2) “We have 90% chance of *observing* at least a 2 mm Hg reduction”
  - E3) “Why did we need such a large trial? So and so’s trial was only half as big and it had a statistically significant result”

- Correct interpretation of 90% power to detect a 2 mm Hg difference:

  “If the true (unknown) treatment effect in population of millions is 2 mm Hg, there is a 90% chance of a statistically significant result (which could happen even if the *observed* treatment effect is < 2 mm Hg)”

Dichotomous Outcomes

- Sometimes outcome is dichotomous
  - Hypertensive at end of study (yes/no)
  - Quit smoking (yes/no)
- Compare treatments using test of proportions (AKA chi-squared test)
- Per-arm sample size for α=.05 two-tailed test with 90% power:

$$n = \frac{\left(1.96\sqrt{2p(1-p)} + 1.28\sqrt{p_T(1-p_T) + p_C(1-p_C)}\right)^2}{\delta^2}$$

- $p=(p_T+p_C)/2$, $\delta=p_T-p_C$ (or p_C-p_T)
- Again for 80% or 85% power, replace 1.28 by .84 or 1.04, respectively

Example

- E.g., trial comparing lifestyle intervention to advice only control for pre-hypertensives
- Outcome: Hypertensive by end of 2 years
- Compare proportion hypertensive in treatment and control arms
- Nuisance parameter p_C

Specifying p_C

- What proportion, p_C, of advice-only patients will develop hypertension in 2 years?
- Want to err on side of underestimating p_C
- Best to use data from other clinical trials in pre-hypertensives, if any
- Often only epi data available

Specifying p_C

- Problems with using epi data:
  - Clinical trial participants may start out healthier than general population (healthy volunteer effect)
  - Clinical trial participants may get better care and may exercise more than general population during trial
- If epi data suggests p_C=.40, might use p_C=.30 to be conservative

Example: Specifying δ

- Usually express treatment effect as percentage reduction, e.g., a 20% relative reduction compared to control
- 30% relative reduction considered large, 10% relative reduction considered small
- Actual number often based on similar trials
- Suppose specify 20% reduction

Specifying δ

- If assume control rate of p_C=.30, a 20% reduction means treatment rate is p_T=(1-.20)(.30)=.24
- (for 15% reduction, p_T=(1-.15)(.30)=.255, etc.)
- Overall rate p=(p_T+p_C)/2=(.30+.24)/2=.27
- Per-arm sample size n is:

$$n = \frac{\left(1.96\sqrt{2(.27)(1-.27)} + 1.28\sqrt{.24(1-.24)+.30(1-.30)}\right)^2}{(.30-.24)^2} = 1148$$

Need 2(1148)=2296 participants total

- Note: Answer is very sensitive to small changes in δ
- Just as in continuous outcome case, halving δ quadruples sample size
- Should do sensitivity analysis:

Power With 1148/Arm

	15% Reduction	20% Reduction	25% Reduction
p_C=.25	.57	.82	.95
p_C=.30	.67	.90	.98
p_C=.35	.77	.95	>.99

- In reality, some people assigned to lifestyle intervention will not comply
- Must take into account when specifying treatment effect; e.g., 20% reduction means 20% net reduction after accounting for fact that some will not comply

- Similarly, some assigned to advice-only may join vigorous workout group
- A 30% advice-only event rate means 30% event rate after accounting for the fact that some will start their own vigorous workout group

- Must also take missing data into account
- For sample size purposes, people often do simple calculation: If need 100/arm to complete study and expect 20% missing data, must randomize 100/.80=125/arm
- Above method not conservative!

Adaptive Sample Size Methods

- Can we look at part of clinical trial data (internal pilot study) and re-compute sample size with revised estimates of nuisance parameters or treatment effect?
- Short answer: No need to worry if based on nuisance parameters, but need to worry if based on treatment effect

- In continuous outcome case, can revise sample size based on std dev estimate with no penalty
- If revise sample size based on data-driven treatment effect estimate, must pay penalty
- In dichotomous outcome case, may revise sample size based on data-driven overall event rate without penalty, but not on treatment effect estimate

Summary

	Continuous Outcome	Dichotomous Outcome
Nuisance Parameter	σ=standard deviation Err on side of overestimating	p_C=control event rate Err on side of underestimating
Treatment Effect	δ=difference in means (for population)	δ=difference in proportions (for population)